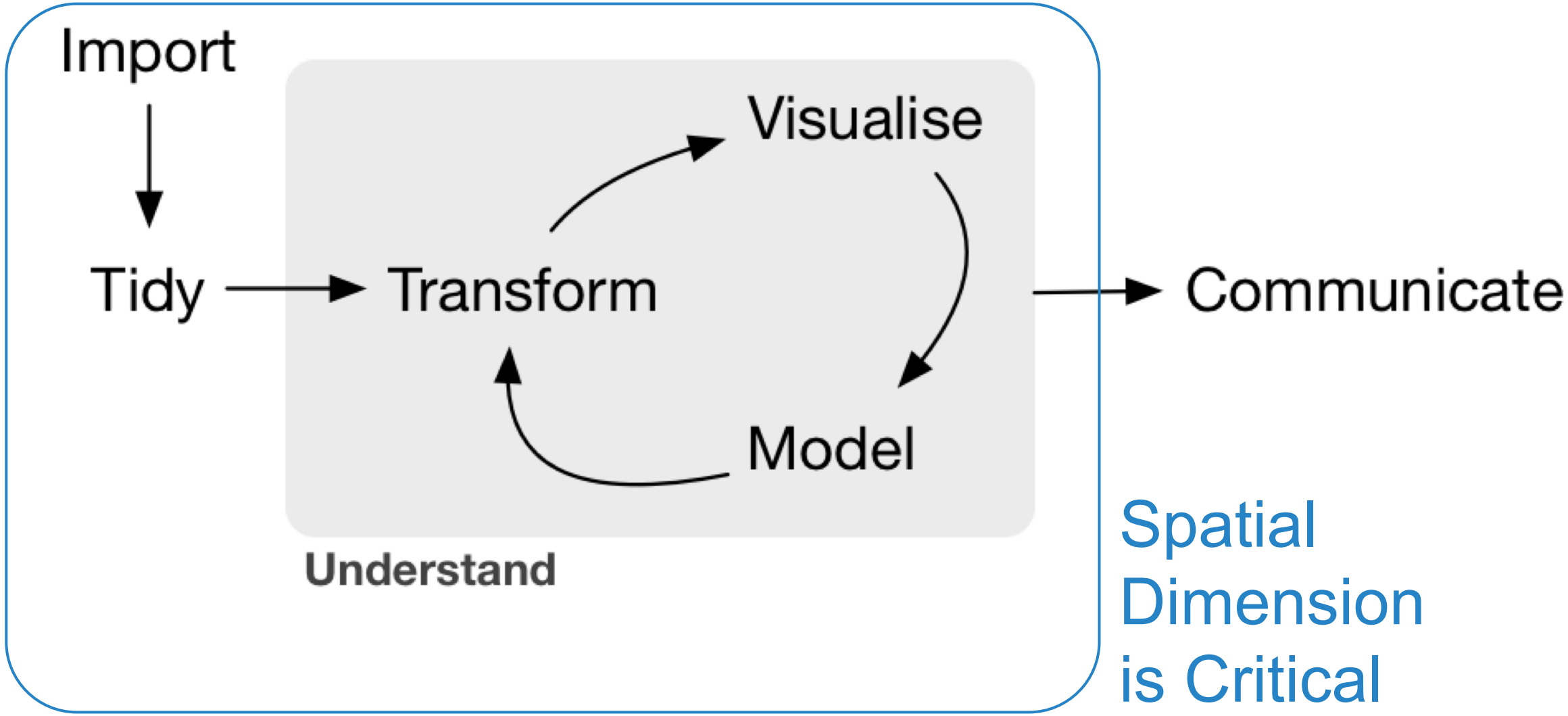
```
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.4.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
```

GEO 503: Spatial Data Science

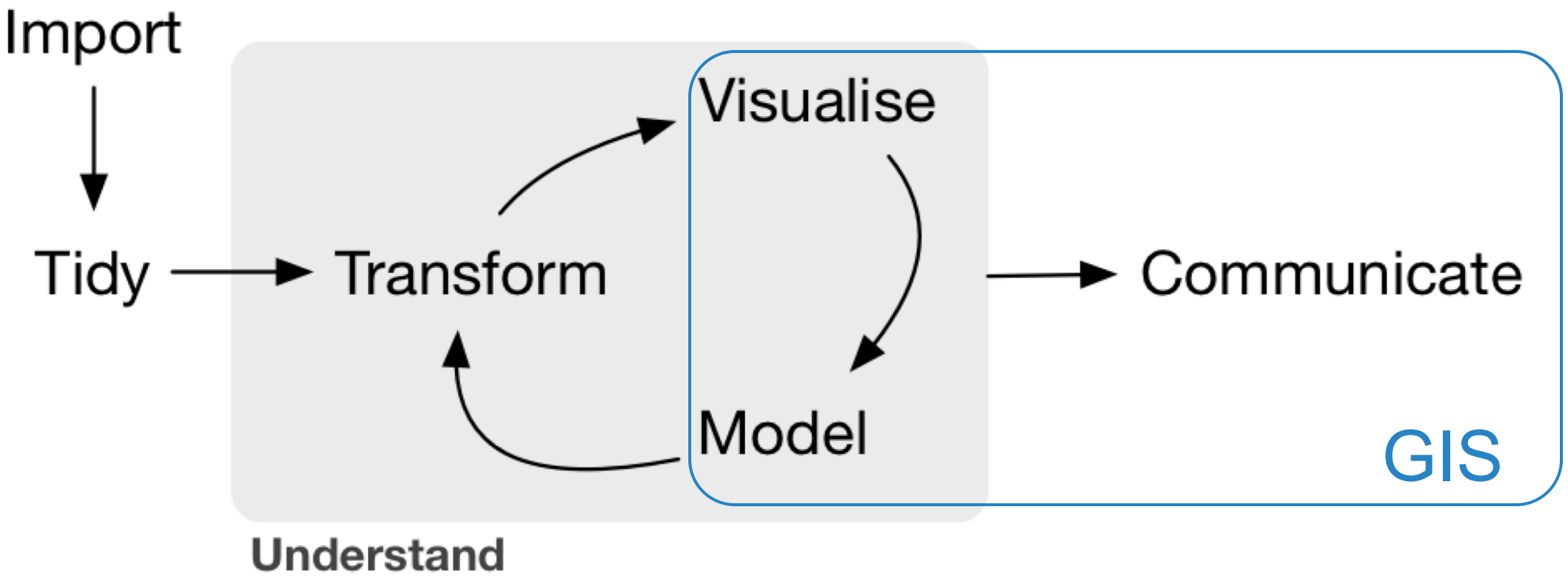
What is data science?

What is Data Science?



Spatial Dimension is Critical

Spatial Data Science vs GIS



Course Website: adamwilson.us/SpatialDataScience

GEO 503: R Spatial Data Science | Home | Syllabus | Schedule | Content | Assignments | Resources

Description

GEO503: R Data Science

Introduction to R for Data Science

Professor Adam M. Wilson
Department of Geography, & Graduate Program in Evolution, Ecology and Behavior
University at Buffalo, Buffalo, NY

Description

The quantity and quality of data available for ecological and environmental research has exploded over the past few decades. These 'big data' now allow us to address important questions (both old and new) with unprecedented rigor and generality. Leveraging these new data streams requires new tools and increasingly sophisticated workflows. The free and open-source R programming language has become a lingua franca for ecological, epidemiological, and statistical research. The course will use a combination of lecture and hands-on exercises to provide a gentle introduction to programming in R with a focus on spatial data processing. The use of 'literate programming' (code embedded within text) to generate dynamic, reproducible research output (figures, manuscripts, websites, etc.) will also be addressed. The course includes an extensive project for students to conduct spatial analysis related to their research. Familiarity with basic GIS concepts (raster, vector,

Course Structure

Course Structure

Tuesdays/Thursdays 2:00-3:20

- Review/Questions
- ~30-45 Minute Presentation
- Guided interactive exercises on your laptops

Course Objectives

4 Learning Objectives

- Become familiar with R programming language
- Learn to code geospatial analyses
- Learn to develop custom data visualization (especially spatial)
- Learn to develop reproducible research workflows

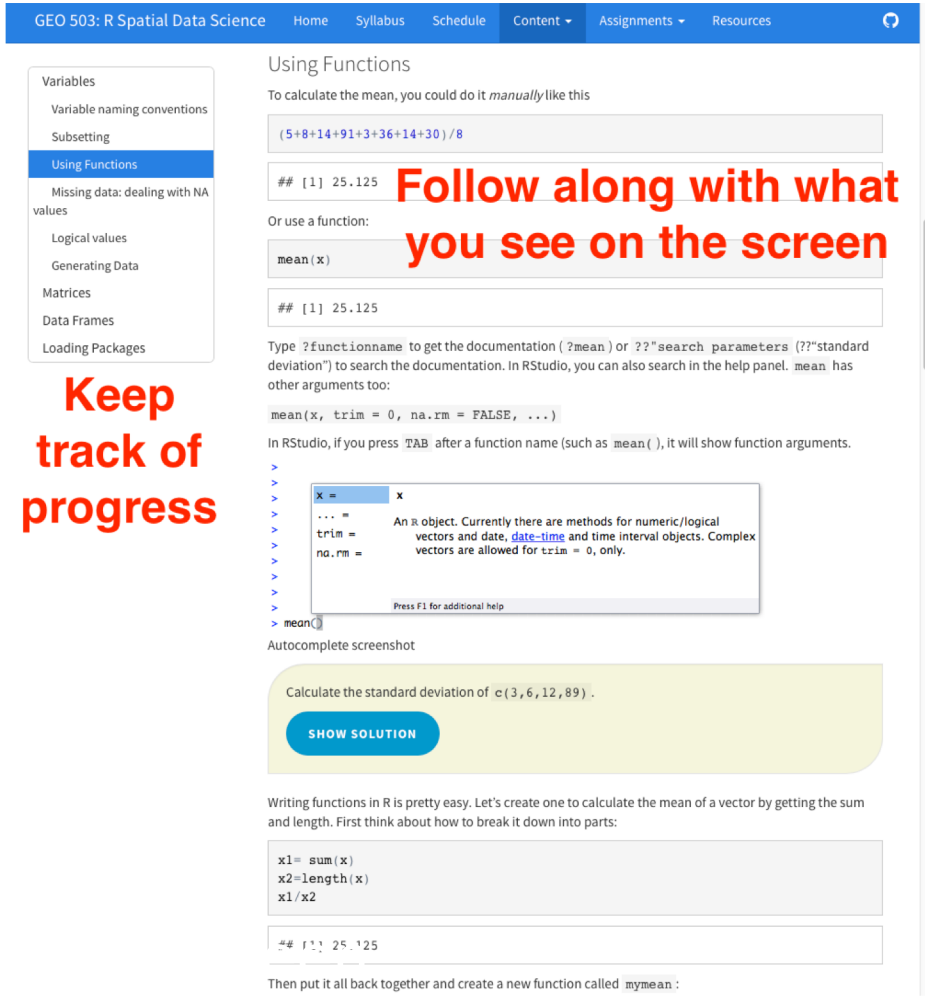
This course is NOT

- A statistics course (see GEO 505, etc.).
- We will focus on workflow and methods

Course Logistics

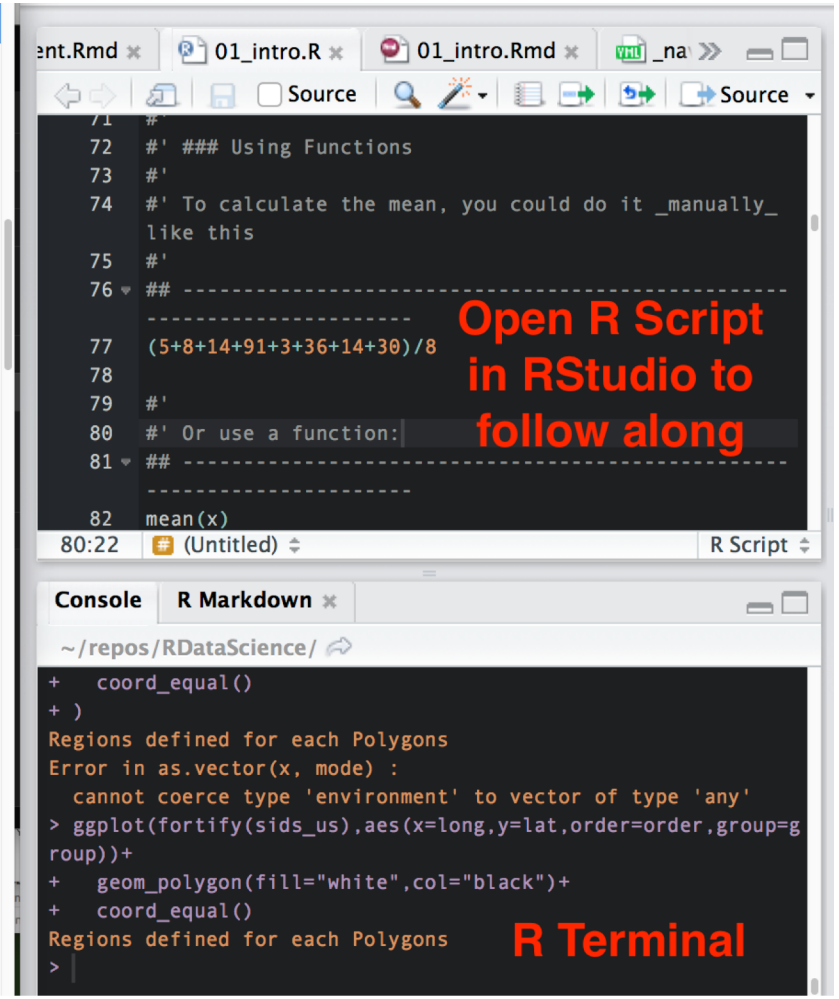
Course Participation	10%
Package Presentation	10%
Homeworks	30%
Final Project	50%

Course Participation (10%)



Follow along with what you see on the screen

Keep track of progress



Open R Script in RStudio to follow along

R Terminal

Package Introduction (10%)

Introduce R package in 5 min presentation

Objectives:

- Learn how to find/download/install a new package and use it
- Teach your peers about useful R packages

The presentation must include:

- What does the package do? (**1-2 slides, 1 minute**)
- Author introduction (**1 slide, 1 minute**)
- Simple demonstration (**2-3 slides, 3 minutes**)

Homework (30%)

```
#' ## Question 1
#' Load the iris dataset by running
## ------------------------------------------------------------------
data(iris)

#'
#' > How many observations (rows) are there for the versicolor species?
#'
#' ____________________
```

Homework submitted in UBlearns

Begin: Homework #1

Cancel Begin

1. Instructions

Description	These quizzes are designed to encourage you to work through the materials we discuss in class *prior* to class so you can come with questions.
Instructions	Please use the attached R script (Homework_01.R) as a template for you to find the answers to the questions. The last question will ask you to upload your updated script (with the code needed to answer the questions). This will not be graded, but will be taken into account if there are any questions about the correct answers later. I recommend that you complete all the questions in the .R file in RStudio before entering the answers into UBLearns.
Force Completion	This test can be saved and resumed later.
Due Date	This Test is due on September 14, 2015 5:00:00 PM EDT.Test cannot be started past this date.

Click **Begin** to start: Homework #1. Click **Cancel** to go back.

2. Submit

Click Begin to start. Click Cancel to quit.

Cancel Begin

Working collaboratively is encouraged but you are responsible for developing your own code to answer the questions:

Acceptable: “which functions did you use to answer #4?”

Unacceptable:“please email me your code for #4.”

Homework format

Take Test: Homework #1

Test Information

Description	These quizzes are designed to encourage you to work through the materials we discuss in class *prior* to class so you can come with questions.
Instructions	Please use the attached R script (Homework_01.R) as a template for you to find the answers to the questions. The last question will ask you to upload your updated script (with the code needed to answer the questions). This will not be graded, but will be taken into account if there are any questions about the correct answers later. I recommend that you complete all the questions in the .R file in RStudio before entering the answers into UBLearns.
Multiple Attempts	Not allowed. This test can only be taken once.
Force Completion	This test can be saved and resumed later.

Question Completion Status:

Save All Answers | Save and Submit

1 points | Save Answer

Question 1

Load the `iris` dataset by running `data(iris)`. How many observations (rows) are there for the versicolor species?

Final Project (50%)

1. Title (<25 words)
2. Introduction [~ 200 words, 10%]
3. Materials and methods [~ 200 words]
   1. Narrative (10%)
   2. Code (25%)
   3. Data (5%)
4. Results [~200 words, 25%]
5. Conclusions [~200 words, 5%]
6. References [5%]

Individual OR small group (≤3) project

"*It takes intelligence, even brilliance, to condense and focus information into a clear, simple presentation that will be read and remembered.*

Ignorance and arrogance are shown in a crowded, complicated, hard-to-read poster."

-- Mary Helen Briscoe

2017 Projects available to browse on website

GEO 503: Spatial Data Science | Home | Syllabus | Schedule | Content | Assignments | Resources

- Homeworks
- Package Introduction
- Final Project
- 2017 Final Projects

2017 Virtual Poster Session

This course uses a combination of lecture and hands-on exercises to provide a gentle introd with a focus on spatial data processing. The final project in the course is the construction of a reproducible researc he figure below.

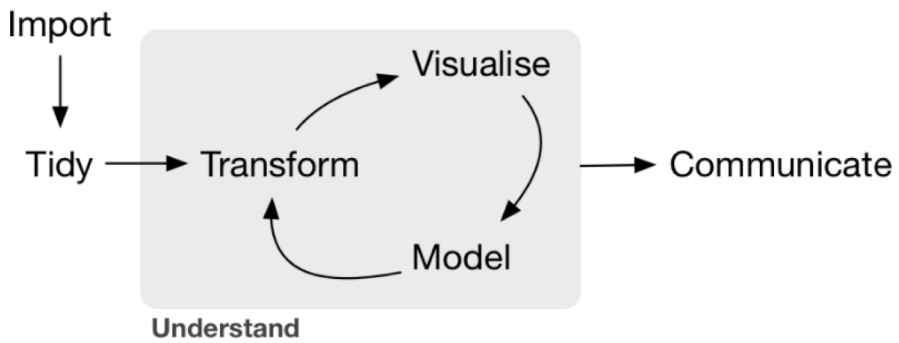


Figure from *R for Data Science* by Grolemund & Wickham (2017)

Each student wrote a script (using the R programming language) to perform these steps and generate a website showcasing their analysis. The focus of the course is on the design and implementation of the complete data processing research workflow itself (not any particular statistics/methods/models). The challenge is to string all the steps together in a *coherent, reproducible flow from raw data to final outputs.*

Student Project Gallery

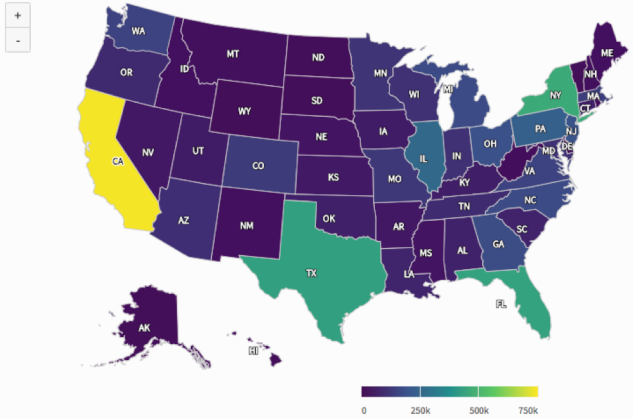
You are invited to explore the student projects below (click on a thumbnail to visit their website). The embedded code reveals their methodological details in addition to their narrative and graphical stories. If you find something interesting, you are free to download and re-run the script to reproduce the entire analysis (from acquiring the original data through generating the tables/figures and even the webpage itself).

Student	Title	Thumbnail
gsaugust	Visualizing Deforestation and Development in Indochina	

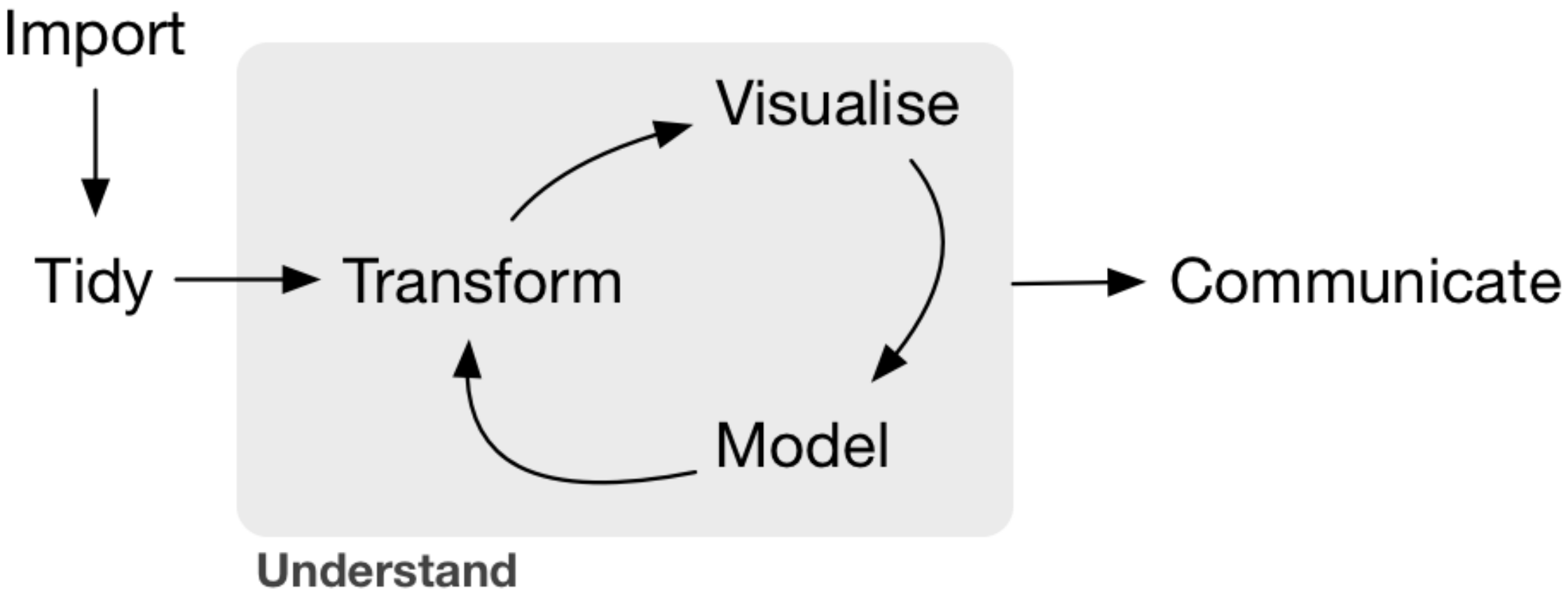
R Data Science Final Project | Home

- Introduction
- Materials and methods
- Results
  - Payroll and Employment
  - Relationship between Payroll and Employment
  - Payroll of interested industries
  - Industries in New York State
  - Spatial Autocorrelation of Information Industry
- Conclusions
- References

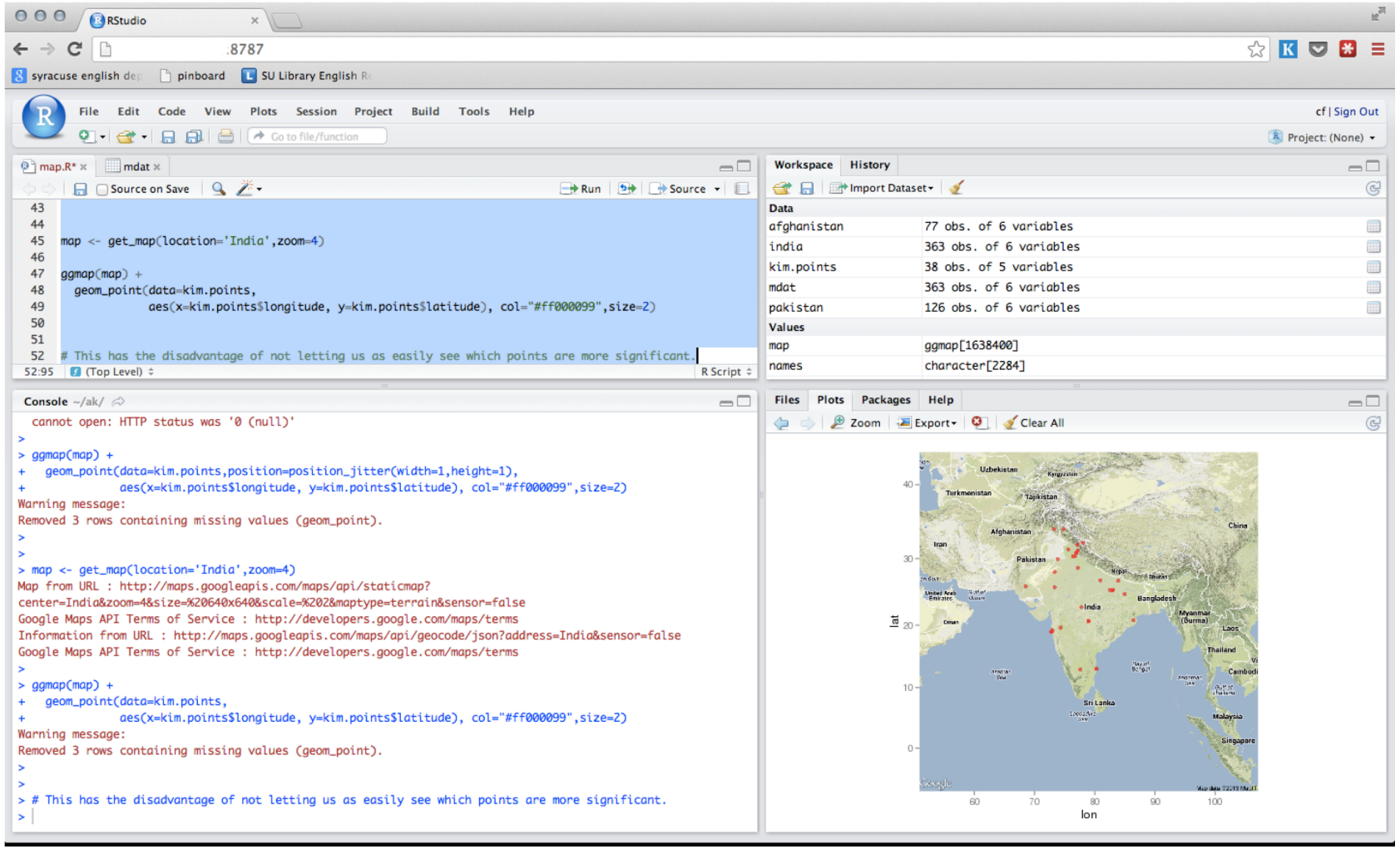
```
    colors = c("#1a6ecc", "#434348", "#90ed7d"),
    chart = list(
      backgroundColor = "transparent",
      style = list(fontFamily = "Source Sans Pro")
    ),
    xAxis = list(
      gridLineWidth = 1
    )
  )
n <- 4
colstops <- data.frame(
  q = 0:n/n,
  c = substring(viridis(n + 1), 0, 7)) %>%
  list_parse2()

highchart() %>%
  hc_add_series_map(usgeojson, subset, name = "Number of Firms",
                    value = "NFirms", joinBy = c("woename", "Geo_Des"),
                    dataLabels = list(enabled = TRUE,
                                      format = '{point.properties.postalcode}')) %>%
  hc_colorAxis(stops = colstops) %>%
  hc_legend(valueDecimals = 0, valueSuffix = "%") %>%
  hc_mapNavigation(enabled = TRUE) %>%
  hc_add_theme(thm)
```



What is Data Science?



Workshop course



Most of our time will be spent thinking about, looking at, and writing code…

Essentially a programming course…

```
adamw — R — 80×21
AdamWilsonMac:~ adamw$ R

R version 3.2.0 (2015-04-16) -- "Full of Ingredients"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.4.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

On the screen

Why Code when you can Click?

Graphical User Interfaces are useful, especially when you are learning...

Reproducible Research

The ability to reproduce results from an experiment or analysis conducted by another*

Developed from literate programming:

- Logic of the analysis is represented in output
- Combines computer code with narrative

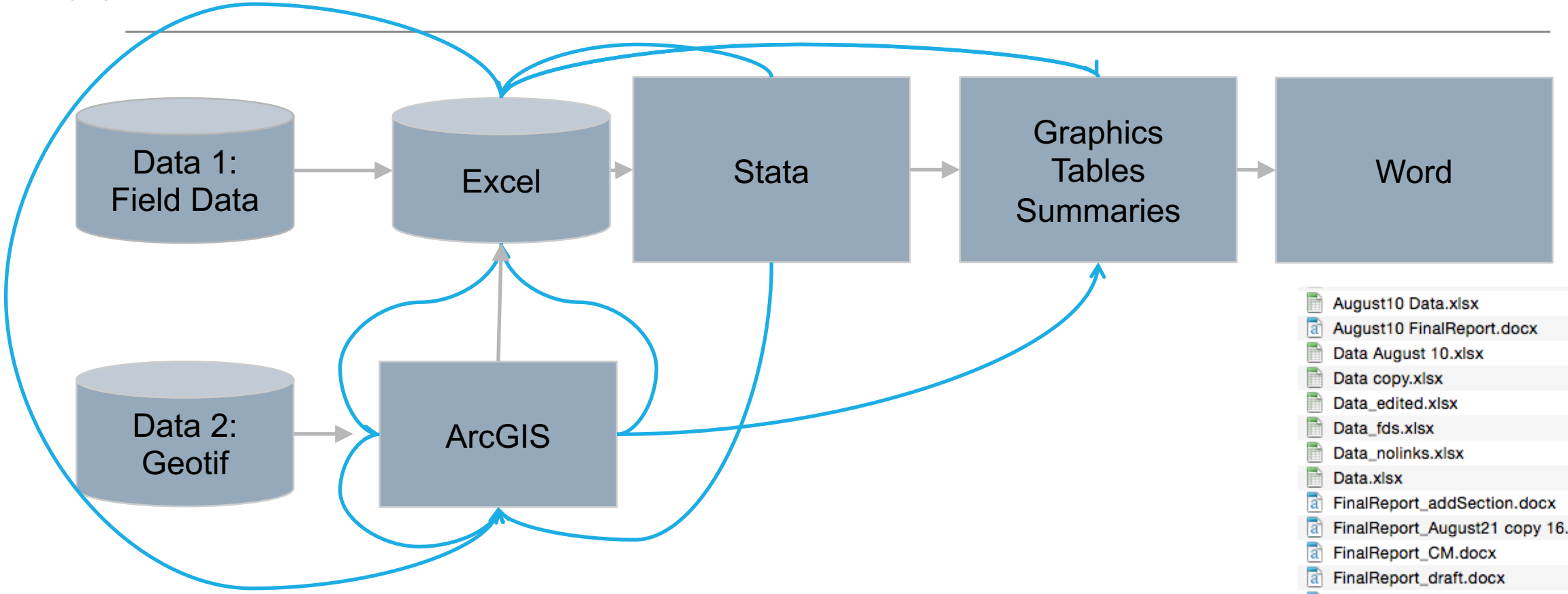
Literate Programming (1992) D.E. Knuth

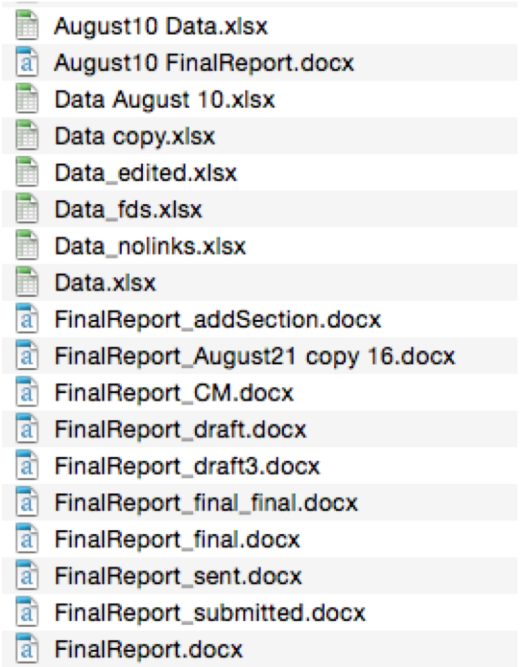
Learning a programming language can help you learn how to think logically.

A man who does not know foreign language is ignorant of his own.

-- Johann Wolfgang von Goethe (1749 - 1832)

Typical GUI Workflow

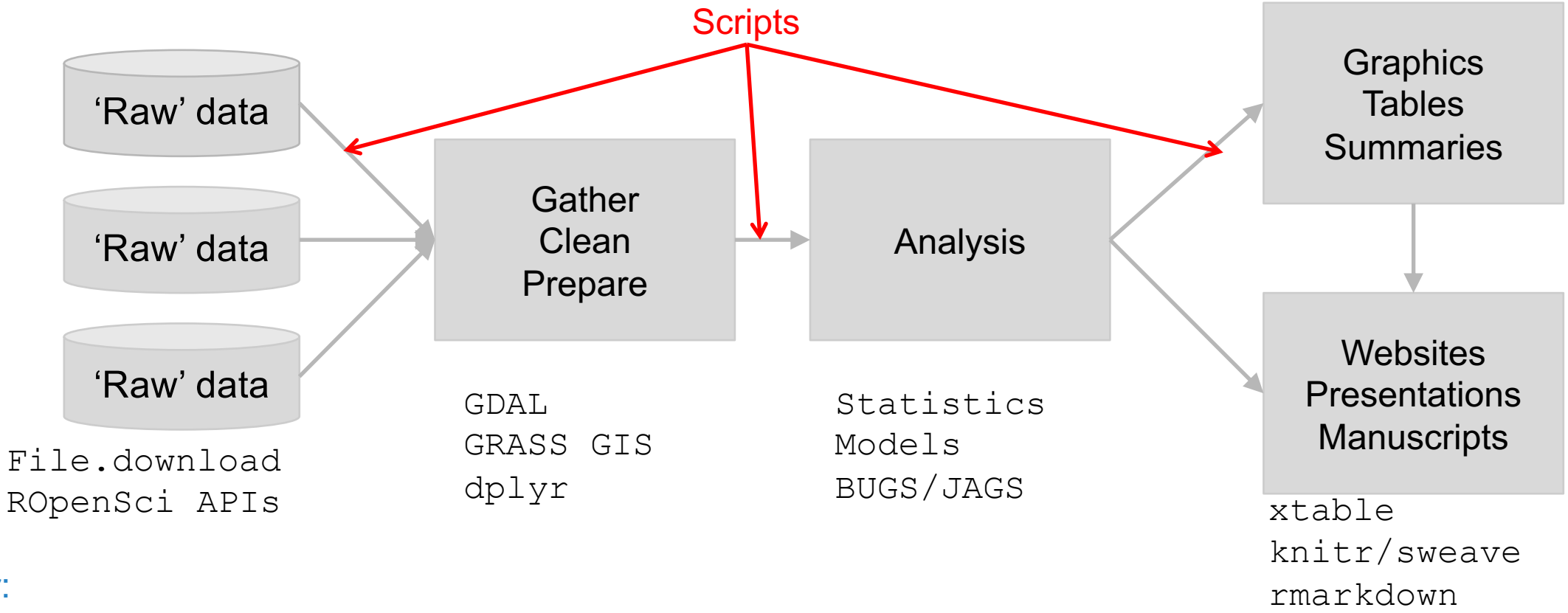




Advisor:

- *I've updated the field data with a few more locations, please re-run that analysis…*
- *New satellite data are available, can you update that figure?*

Organized and repeatable workflow (and some example commands)



Advisor:

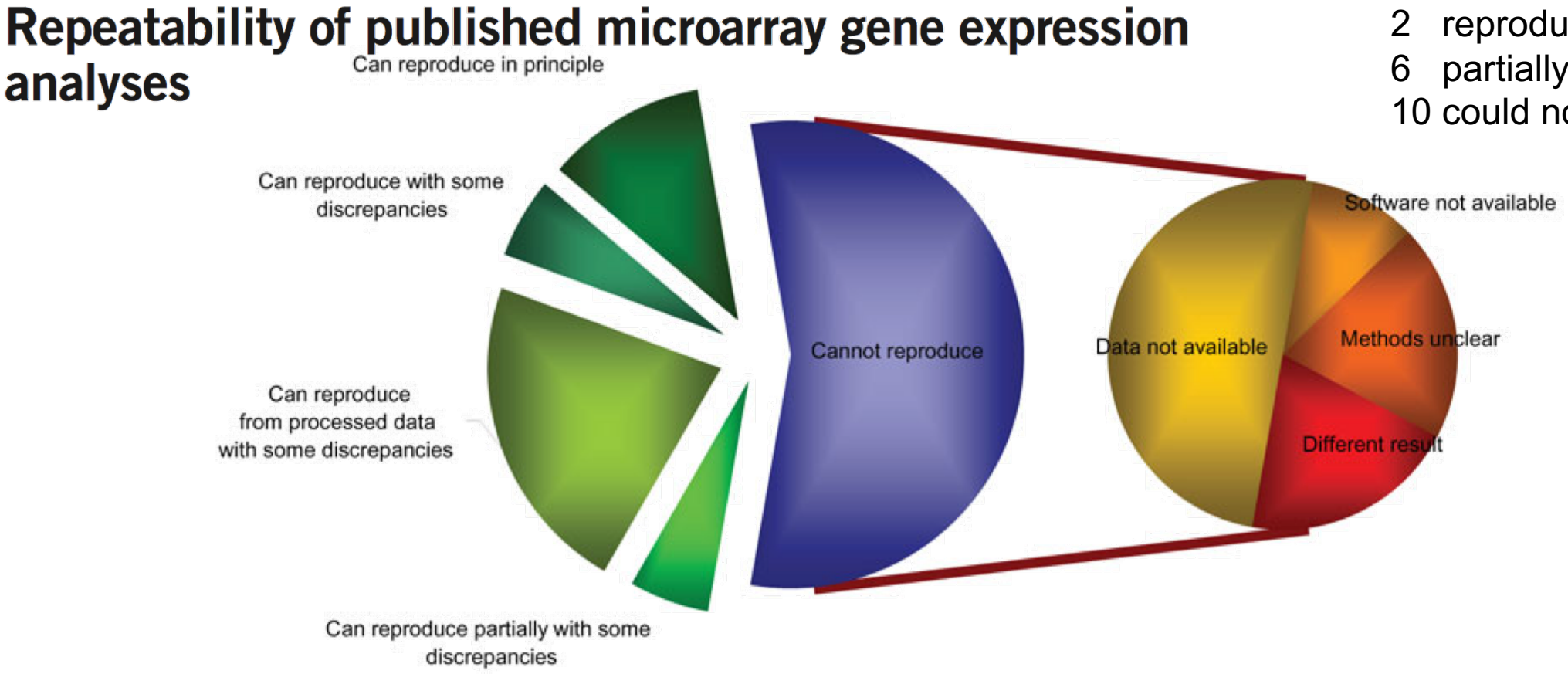
- *I've updated the field data with a few more locations, please re-run that analysis…*
- *New satellite data are available, can you update that figure?*

Sure, I can do that this afternoon…

Adapted from Gandrud (2014) *Reproducible Research with R and RStudio*.

Reproducible Research

Repeatability of published microarray gene expression Analyses. *Nature Genetics* 41(2):149



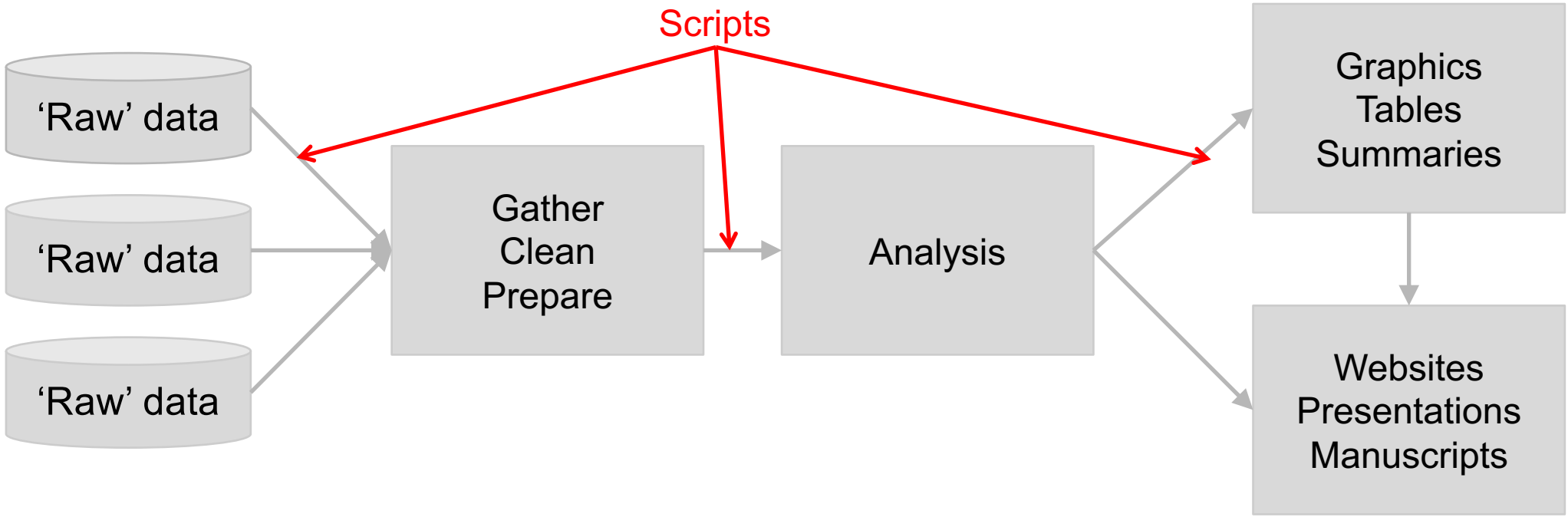
18 articles:
- 2 reproducible
- 6 partially
- 10 could not be reproduced!

Programming gives you access to more computer power.

The computer is incredibly fast, accurate, and stupid. Man is unbelievably slow, inaccurate, and brilliant. The marriage of the two is a force beyond calculation.

-- Leo Cherne

Organized and repeatable workflow (and some example commands)

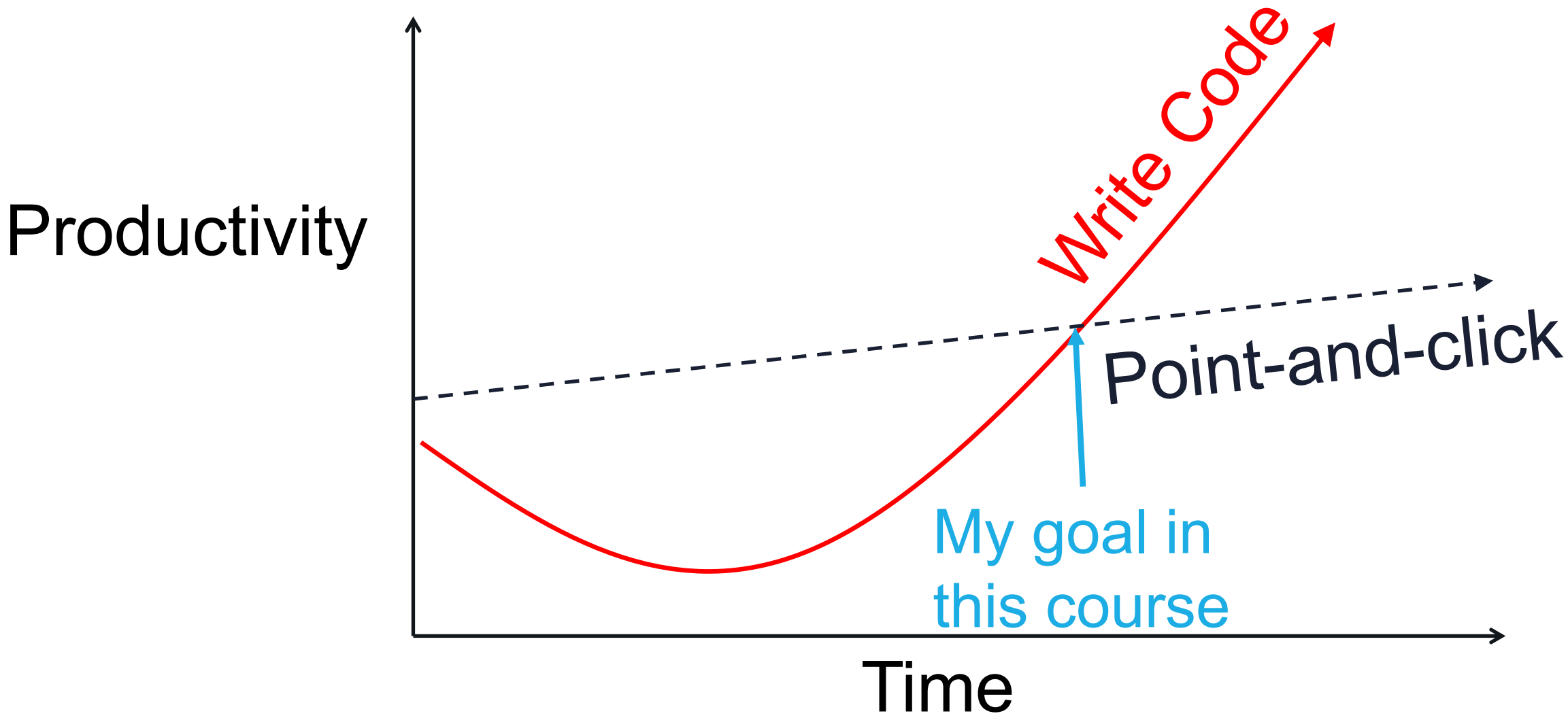


Advisor:

- *I want you to take he analysis you developed for Buffalo and run it globally*

Sure, I can do that this afternoon…

From Graphical User Interface (GUI) to scripting/programming



Typical software use - GEO

Software

- ArcGIS 94%
- Python 29%
- R 29%
- SPSS 29%
- Google Earth Engine 24%
- Erdas Imagine 24%

Scripting

- Yes 71%
- No 29%

Used R?

- No 52%

The R Project for Statistical Computing

Free and Open source

Data manipulation

Data analysis tools

Great graphics

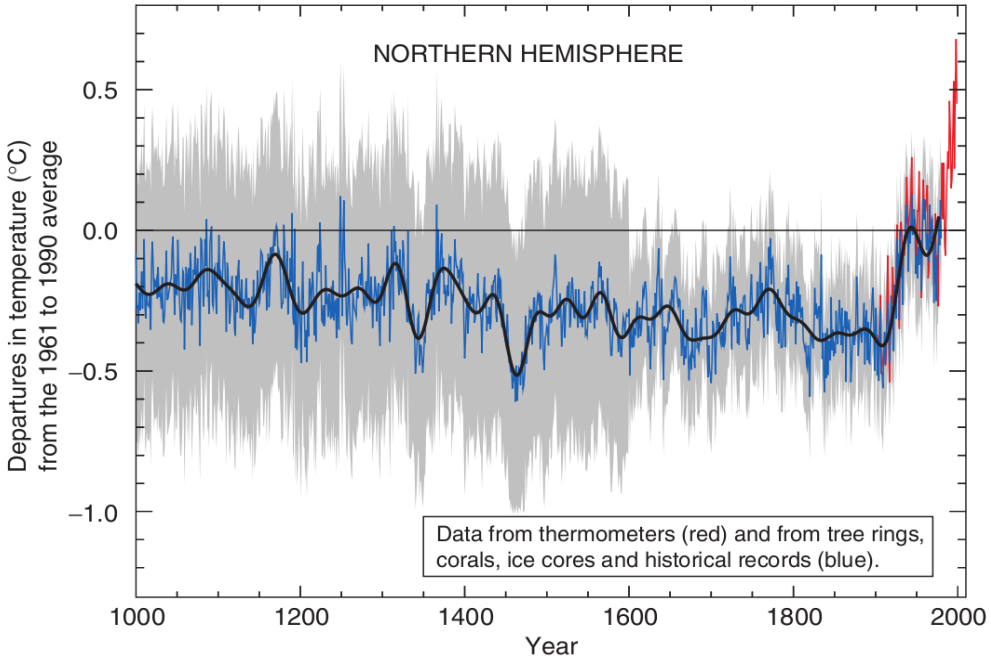
Programming language

6,000^{+} free, community-contributed packages

A supportive and increasing user community

R is a dialect of the S language developed at Bell Laboratories (formerly AT&T) by John Chambers et. al. (same group developed C and UNIX©)

Reproducible, Portable, & Transparent



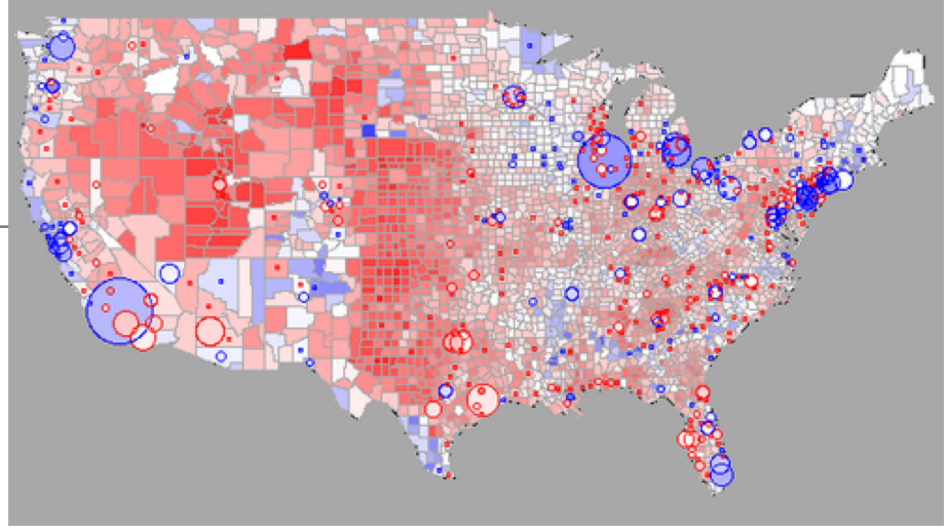
. . . all the code and data used to recreate the Mann's original analysis has been made available to the public [...] Since the analysis is in R, anyone can replicate the results and examine the methods.

(Matthew Pocernich, R news 6/4, 10/31/06)

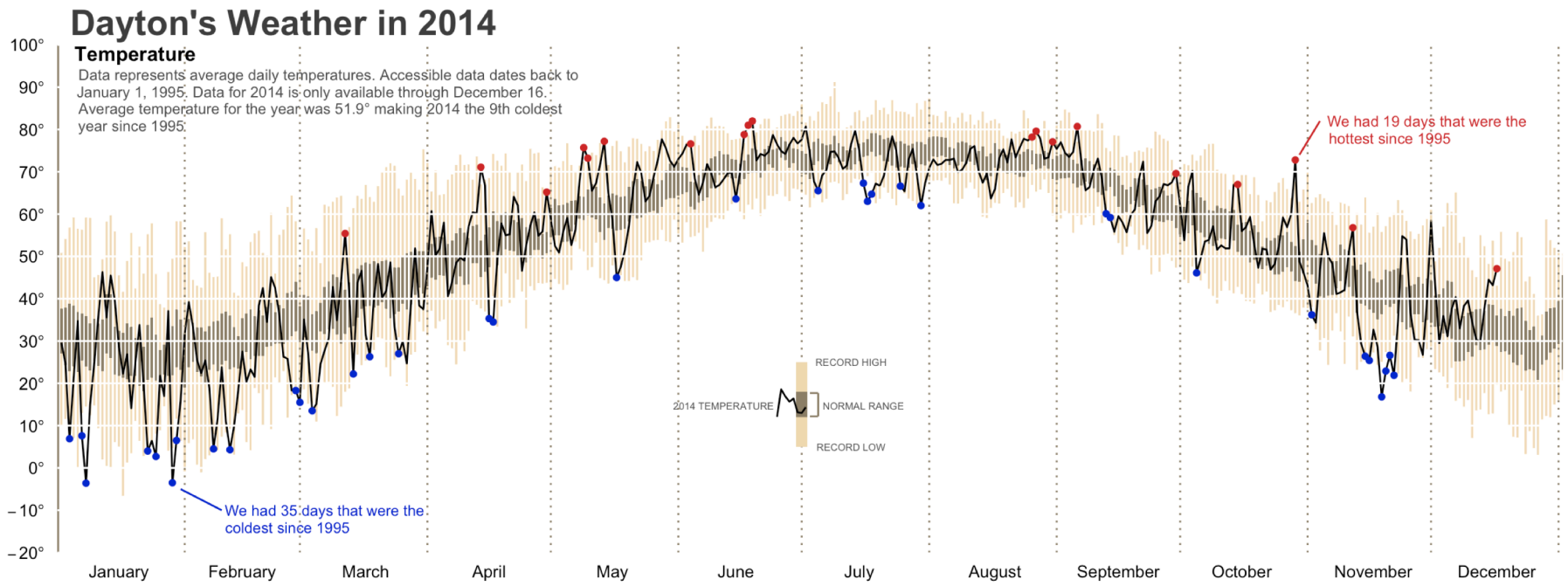
http://www.cgd.ucar.edu/ccr/ammann/millennium/refs/WahlAmmann_ClimChange2006.html

R Graphics

If you can imagine it…



http://blog.revolutionanalytics.com/2009/01/r-graph-gallery.html



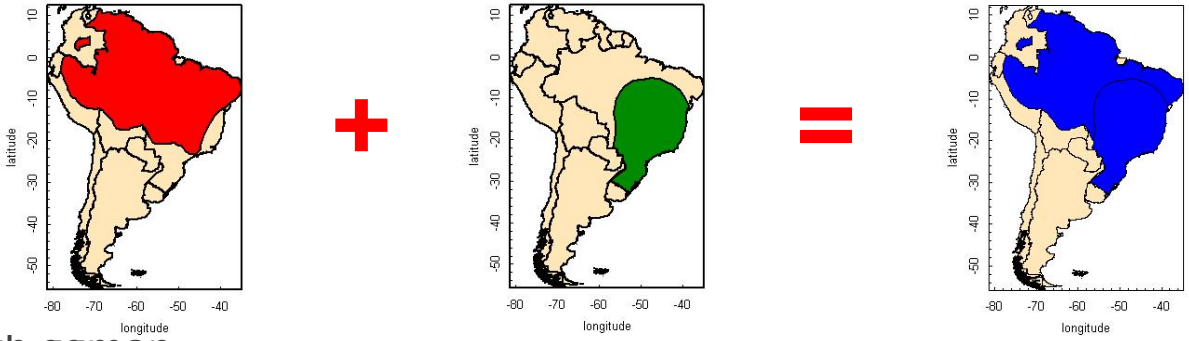
http://rpubs.com/bradleyboehmke/weather_graphic

Spatial data in R

Packages: sp, maptools, rgeos, raster, ggmap

Examples:

- species range overlays



- Basemaps with ggmap

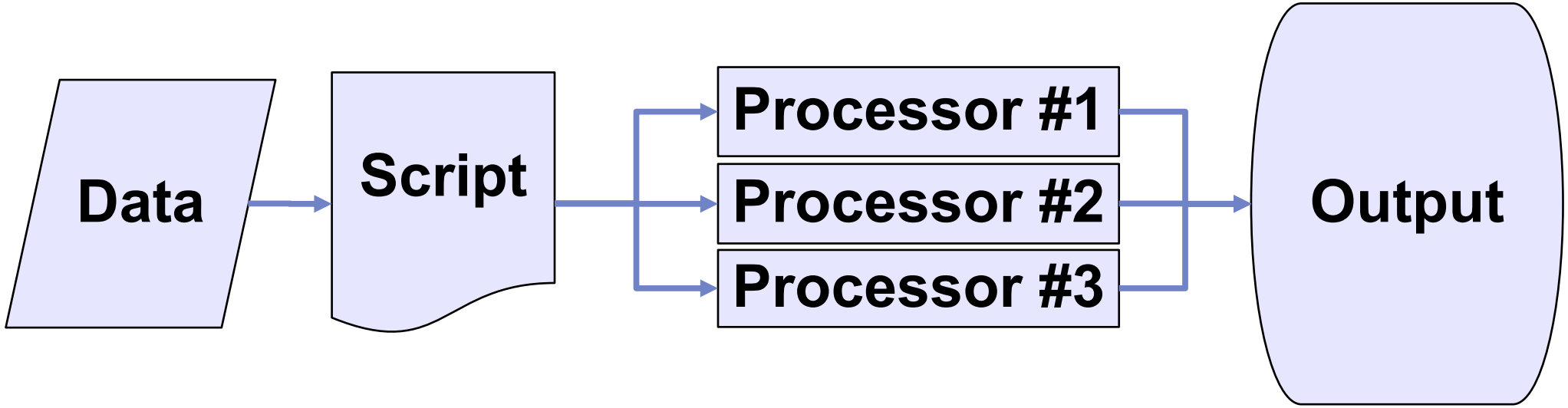
http://www.nceas.ucsb.edu/

http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf

Parallel Processing

For BIG jobs:

multi-core processors / high performance computing with foreach.



Strengths & Limitations

Just-in-time compilation:

- Slower than compiled languages (-)
- Faster to compose (+)
- Many available packages (+)

Most operations conducted in RAM

- RAM can be limiting and/or expensive (-)
  - "Error: cannot allocate vector of size *X* Mb"
- Various packages and clever programming can overcome this… (+)

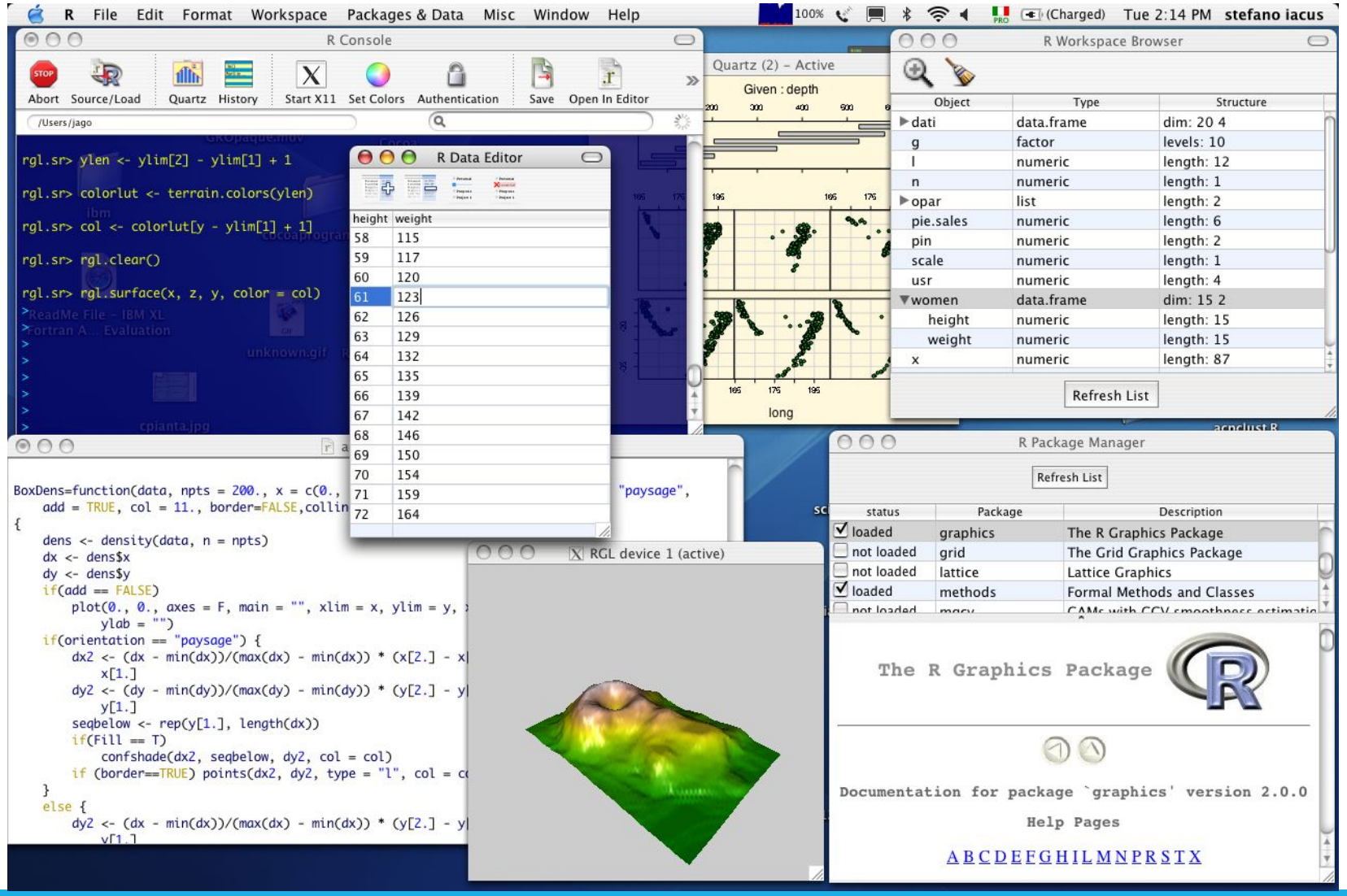
Free like beer and speech! (+)

R Interface

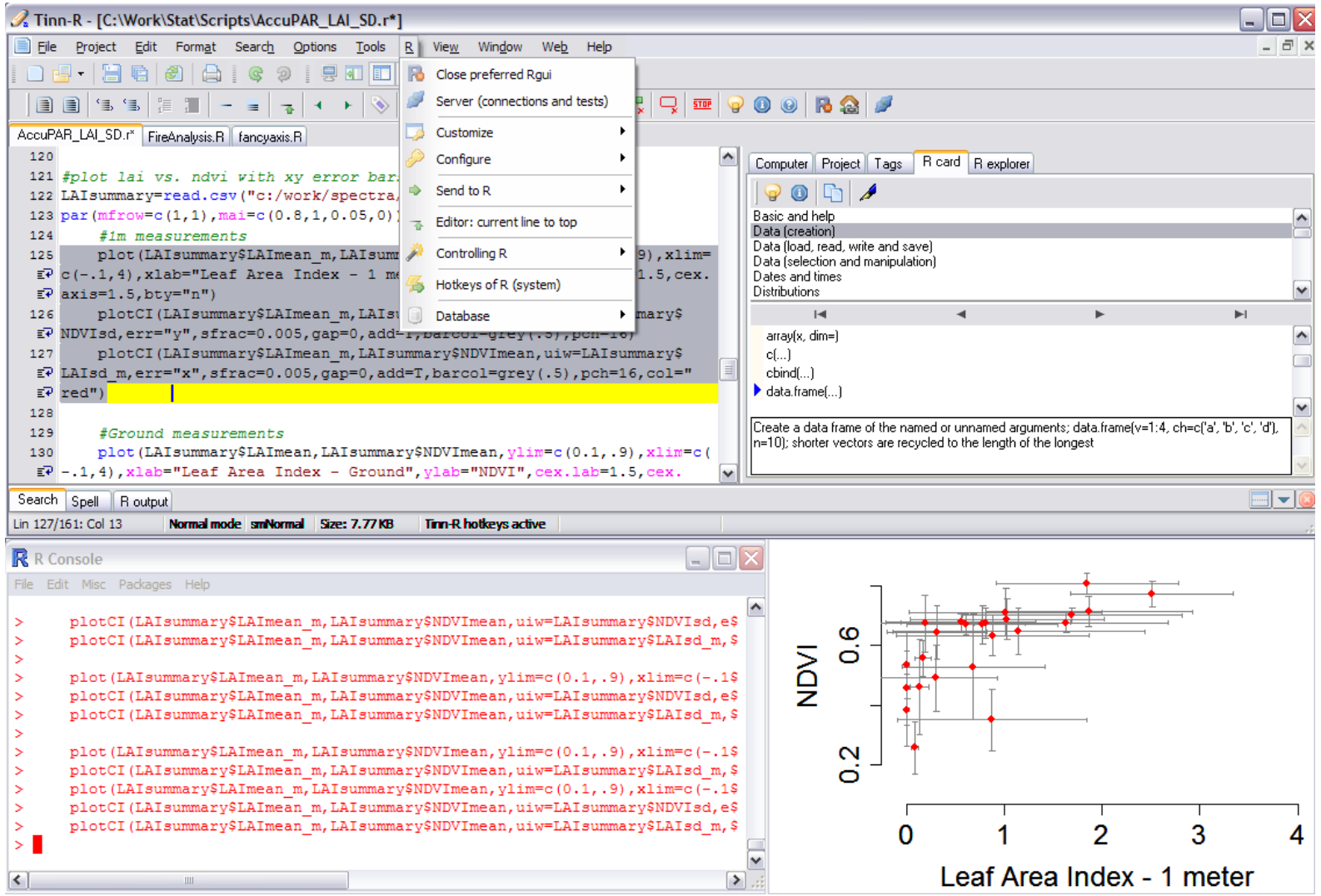
```
adamw — R — 80×21
AdamWilsonMac:~ adamw$ R

R version 3.2.0 (2015-04-16) -- "Full of Ingredients"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.4.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

But there are other options...

R in Mac


R File Edit Format Workspace Packages & Data Misc Window Help
Tue 2:14 PM stefano iacus
R Console
Abort Source/Load Quartz History Start X11 Set Colors Authentication Save Open In Editor
/Users/jago
rgl.sr> ylen <- ylim[2] - ylim[1] + 1
rgl.sr> colorlut <- terrain.colors(ylen)
rgl.sr> col <- colorlut[y - ylim[1] + 1]
rgl.sr> rgl.clear()
rgl.sr> rgl.surface(x, z, y, color = col)
R Data Editor
height weight
58 115
59 117
60 120
61 123
62 126
63 129
64 132
65 135
66 139
67 142
68 146
69 150
70 154
71 159
72 164
Quartz (2) - Active
Given : depth
long
R Workspace Browser
Object Type Structure
dati data.frame dim: 20 4
g factor levels: 10
l numeric length: 12
n numeric length: 1
opar list length: 2
pie.sales numeric length: 6
pin numeric length: 2
scale numeric length: 1
usr numeric length: 4
women data.frame dim: 15 2
height numeric length: 15
weight numeric length: 15
x numeric length: 87
Refresh List
BoxDens=function(data, npts = 200., x = c(0.,
add = TRUE, col = 11., border=FALSE,collin
"paysage",
dens <- density(data, n = npts)
dx <- dens$x
dy <- dens$y
if(add == FALSE)
plot(0., 0., axes = F, main = "", xlim = x, ylim = y,
ylab = "")
if(orientation == "paysage") {
dx2 <- (dx - min(dx))/(max(dx) - min(dx)) * (x[2.] - x
x[1.]
dy2 <- (dy - min(dy))/(max(dy) - min(dy)) * (y[2.] - y
y[1.]
seqbelow <- rep(y[1.], length(dx))
if(Fill == T)
confshade(dx2, seqbelow, dy2, col = col)
if (border==TRUE) points(dx2, dy2, type = "l", col = c
}
else {
dy2 <- (dx - min(dx))/(max(dx) - min(dx)) * (y[2.] - y
RGL device 1 (active)
R Package Manager
Refresh List
status Package Description
loaded graphics The R Graphics Package
not loaded grid The Grid Graphics Package
not loaded lattice Lattice Graphics
loaded methods Formal Methods and Classes
The R Graphics Package
Documentation for package `graphics' version 2.0.0
Help Pages
A B C D E F G H I L M N P R S T X

R in Windows

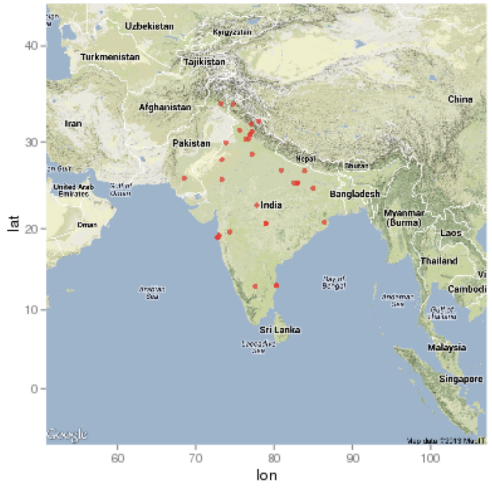

Tinn-R - [C:\Work\Stat\Scripts\AccuPAR_LAI_SD.r*]
File Project Edit Format Search Options Tools R View Window Web Help
Close preferred Rgui
Server (connections and tests)
Customize
Configure
Send to R
Editor: current line to top
Controlling R
Hotkeys of R (system)
Database
AccuPAR_LAI_SD.r* FireAnalysis.R fancyaxis.R
#plot lai vs. ndvi with xy error bar
LAIsummary=read.csv("c:/work/spectra/
par(mfrow=c(1,1),mai=c(0.8,1,0.05,0)
#1m measurements
plot(LAIsummary$LAImean_m,LAIsum
c(-.1,4),xlab="Leaf Area Index - 1 me
axis=1.5,bty="n")
plotCI(LAIsummary$LAImean_m,LAIsu
NDVIsd,err="y",sfrac=0.005,gap=0,add=T,barcol=grey(.5),pch=16)
plotCI(LAIsummary$LAImean_m,LAIsummary$NDVImean,uiw=LAIsummary$
LAIsd_m,err="x",sfrac=0.005,gap=0,add=T,barcol=grey(.5),pch=16,col="
red")
#Ground measurements
plot(LAIsummary$LAImean,LAIsummary$NDVImean,ylim=c(0.1,.9),xlim=c(
-.1,4),xlab="Leaf Area Index - Ground",ylab="NDVI",cex.lab=1.5,cex.
Computer Project Tags R card R explorer
Basic and help
Data (creation)
Data (load, read, write and save)
Data (selection and manipulation)
Dates and times
Distributions
array(x, dim=)
c(...)
cbind(...)
data.frame(...)
Create a data frame of the named or unnamed arguments; data.frame(v=1:4, ch=c('a', 'b', 'c', 'd'), n=10); shorter vectors are recycled to the length of the longest
Search Spell R output
Lin 127/161: Col 13 Normal mode smNormal Size: 7.77 KB Tinn-R hotkeys active
R Console
File Edit Misc Packages Help
> plotCI(LAIsummary$LAImean_m,LAIsummary$NDVImean,uiw=LAIsummary$NDVIsd,e$
> plotCI(LAIsummary$LAImean_m,LAIsummary$NDVImean,uiw=LAIsummary$LAIsd_m,$
> plot(LAIsummary$LAImean_m,LAIsummary$NDVImean,ylim=c(0.1,.9),xlim=c(-.1$
> plotCI(LAIsummary$LAImean_m,LAIsummary$NDVImean,uiw=LAIsummary$NDVIsd,e$
> plotCI(LAIsummary$LAImean_m,LAIsummary$NDVImean,uiw=LAIsummary$LAIsd_m,$
> plot(LAIsummary$LAImean_m,LAIsummary$NDVImean,ylim=c(0.1,.9),xlim=c(-.1$
> plotCI(LAIsummary$LAImean_m,LAIsummary$NDVImean,uiw=LAIsummary$LAIsd_m,$
> plot(LAIsummary$LAImean_m,LAIsummary$NDVImean,ylim=c(0.1,.9),xlim=c(-.1$
> plotCI(LAIsummary$LAImean_m,LAIsummary$NDVImean,uiw=LAIsummary$NDVIsd,e$
> plotCI(LAIsummary$LAImean_m,LAIsummary$NDVImean,uiw=LAIsummary$LAIsd_m,$
NDVI
0.2
0.6
0
1
2
3
4
Leaf Area Index - 1 meter

RStudio

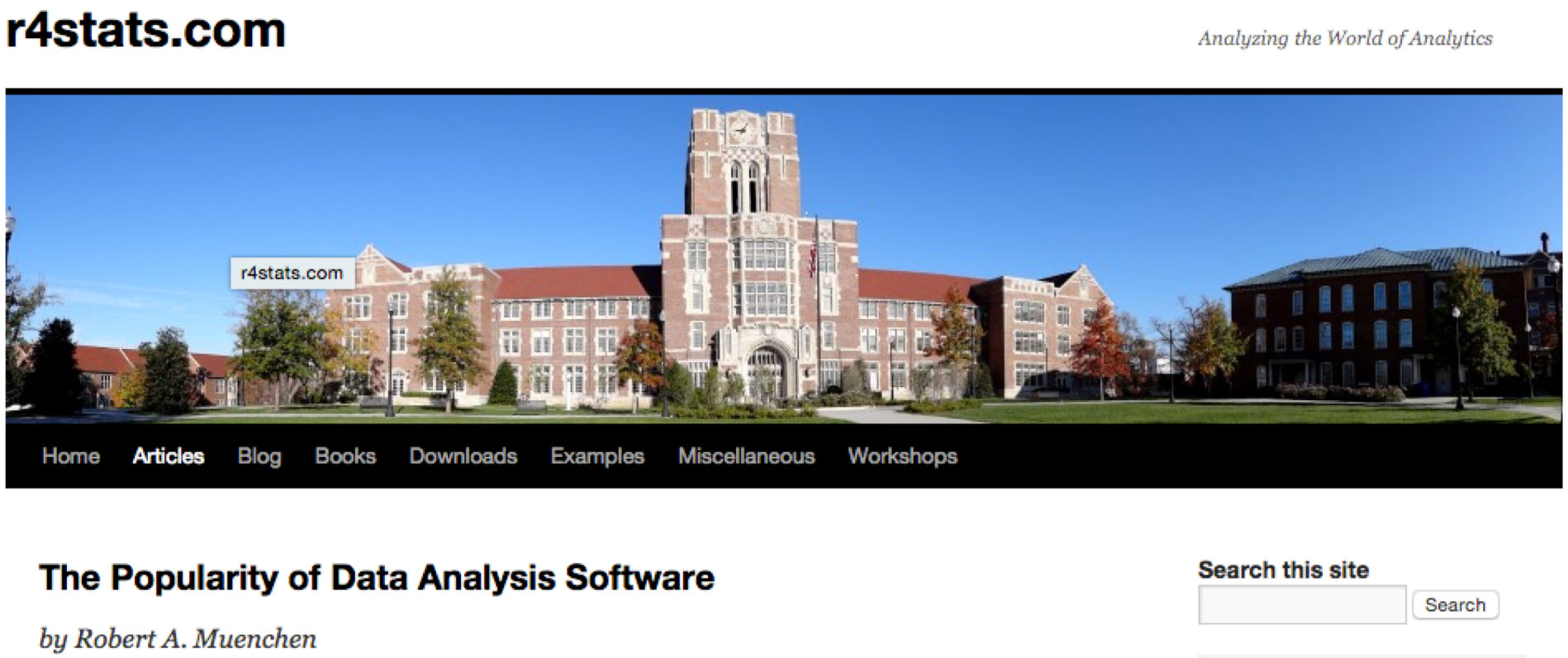
```
map <- get_map(location='India',zoom=4)

ggmap(map) +
  geom_point(data=kim.points,
             aes(x=kim.points$longitude, y=kim.points$latitude), col="#ff000099",size=2)

# This has the disadvantage of not letting us as easily see which points are more significant.
```

```
cannot open: HTTP status was '0 (null)'
>
> ggmap(map) +
+   geom_point(data=kim.points,position=position_jitter(width=1,height=1),
+              aes(x=kim.points$longitude, y=kim.points$latitude), col="#ff000099",size=2)
Warning message:
Removed 3 rows containing missing values (geom_point).
>
>
> map <- get_map(location='India',zoom=4)
Map from URL : http://maps.googleapis.com/maps/api/staticmap?
center=India&zoom=4&size=%20640x640&scale=%202&maptype=terrain&sensor=false
Google Maps API Terms of Service : http://developers.google.com/maps/terms
Information from URL : http://maps.googleapis.com/maps/api/geocode/json?address=India&sensor=false
Google Maps API Terms of Service : http://developers.google.com/maps/terms
>
> ggmap(map) +
+   geom_point(data=kim.points,
+              aes(x=kim.points$longitude, y=kim.points$latitude), col="#ff000099",size=2)
Warning message:
Removed 3 rows containing missing values (geom_point).
>
>
> # This has the disadvantage of not letting us as easily see which points are more significant.
>
```

Data	
afghanistan	77 obs. of 6 variables
india	363 obs. of 6 variables
kim.points	38 obs. of 5 variables
mdat	363 obs. of 6 variables
pakistan	126 obs. of 6 variables
Values	
map	ggmap[1638400]
names	character[2284]

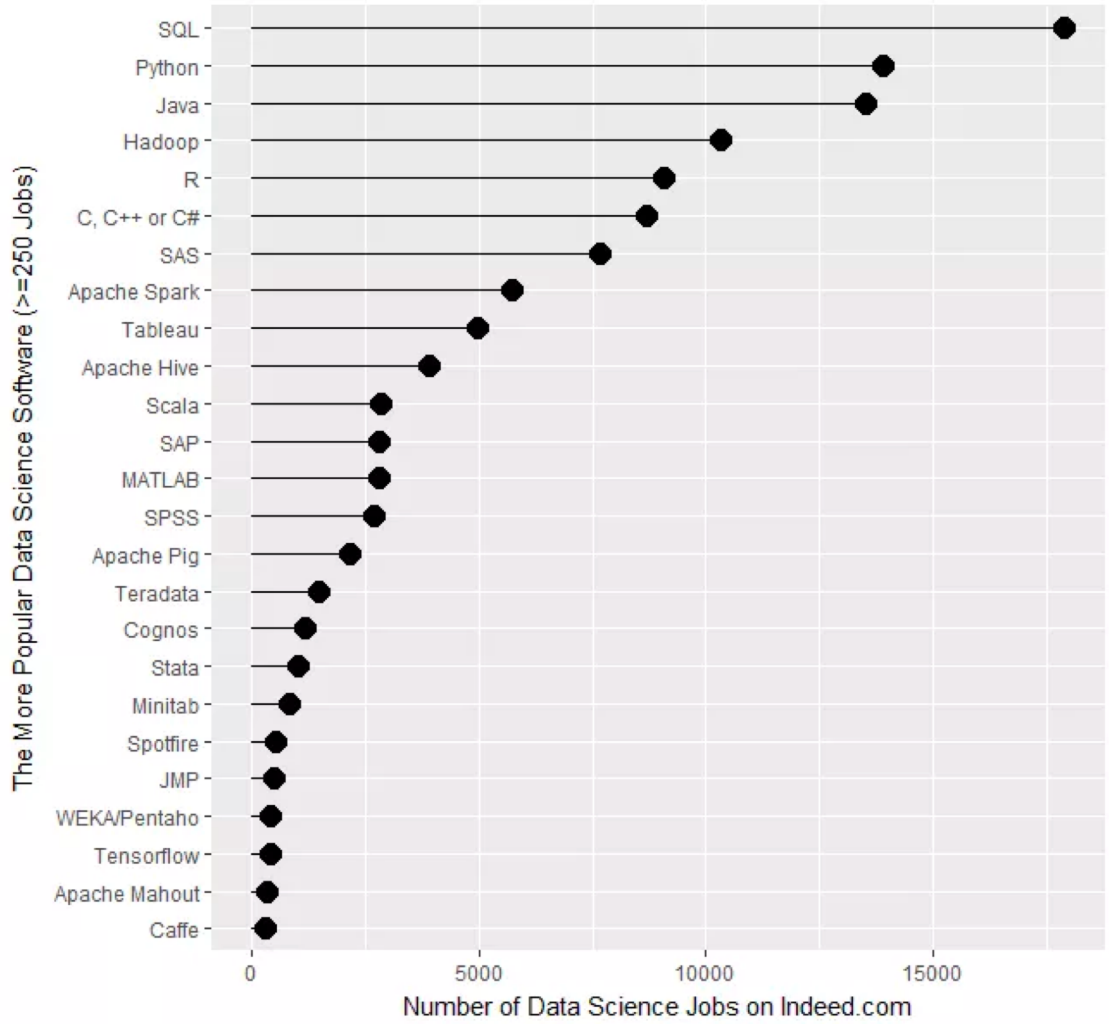


Who uses R?



http://r4stats.com/articles/popularity/

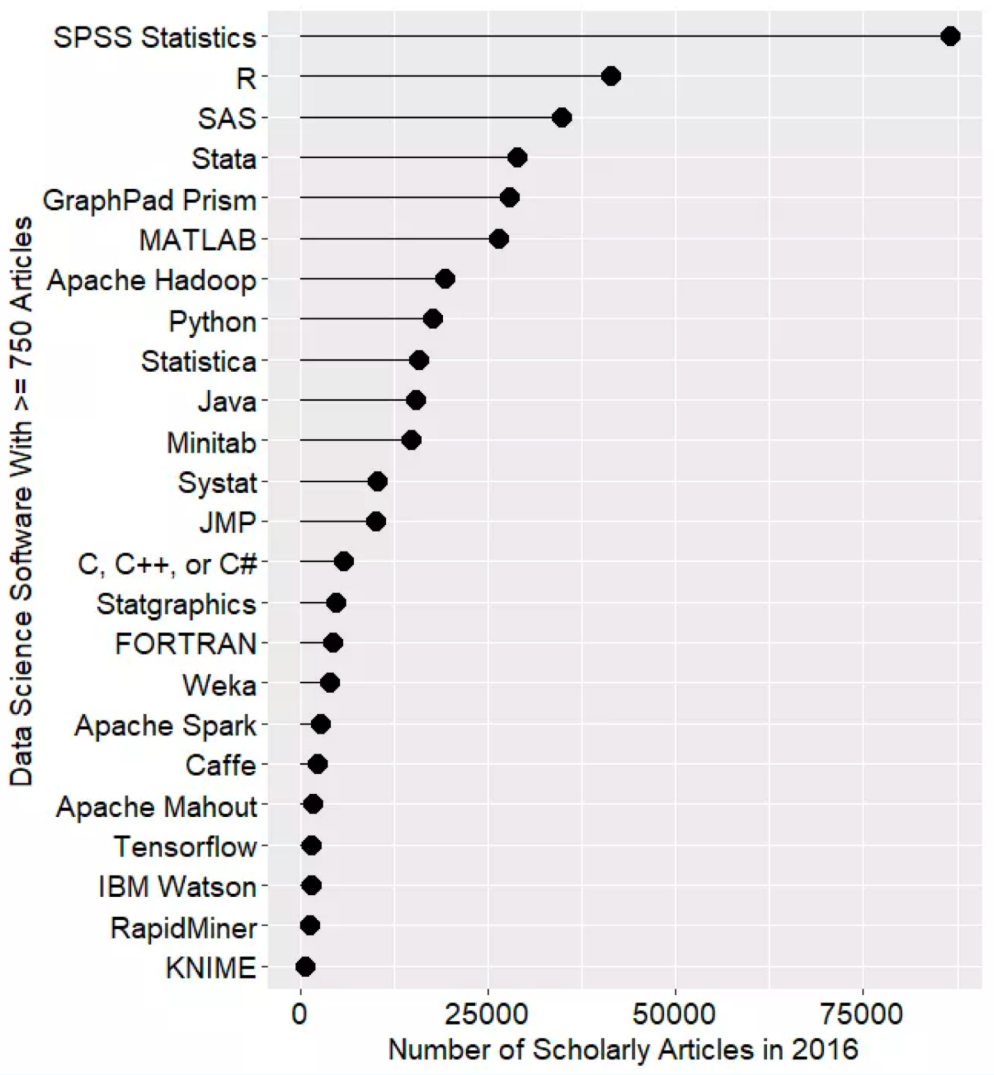
“Data Science” Jobs on indeed.com



The number of data science jobs for the more popular software (those with 250 jobs or more, 2/2017).

Feb 2017: http://r4stats.com/articles/popularity/

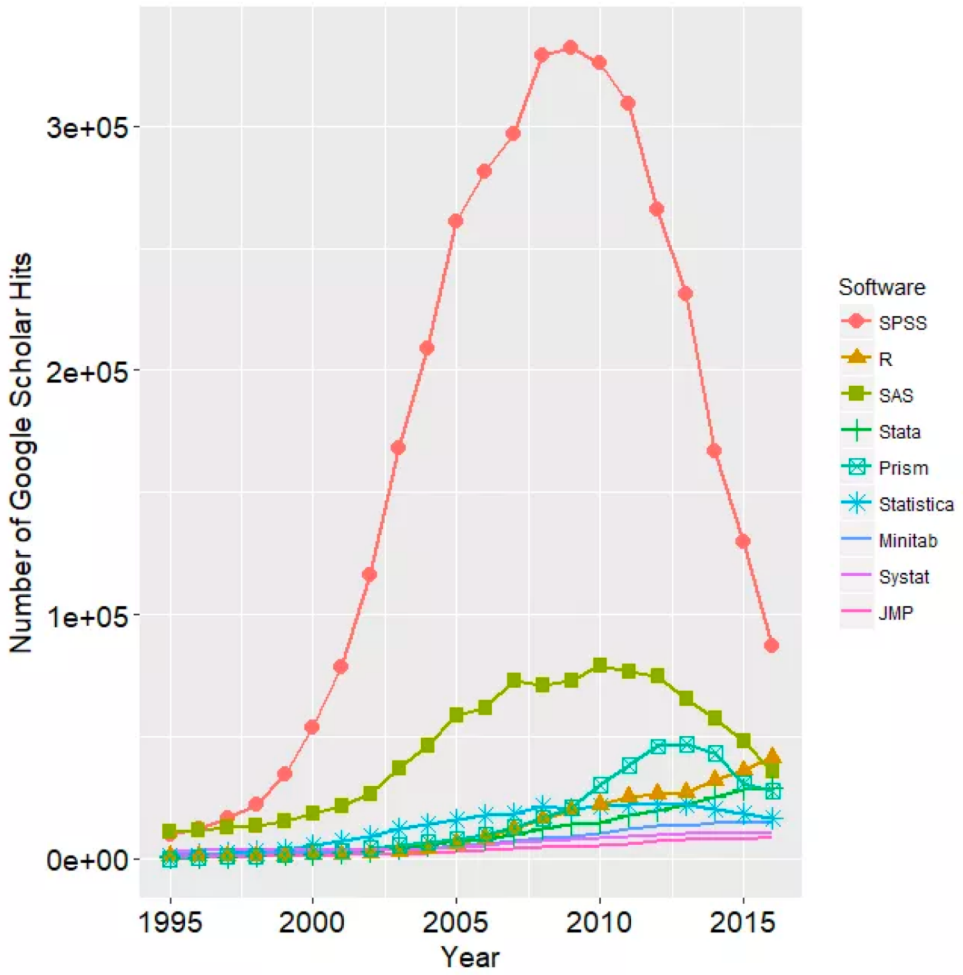
Scholarly articles by software package



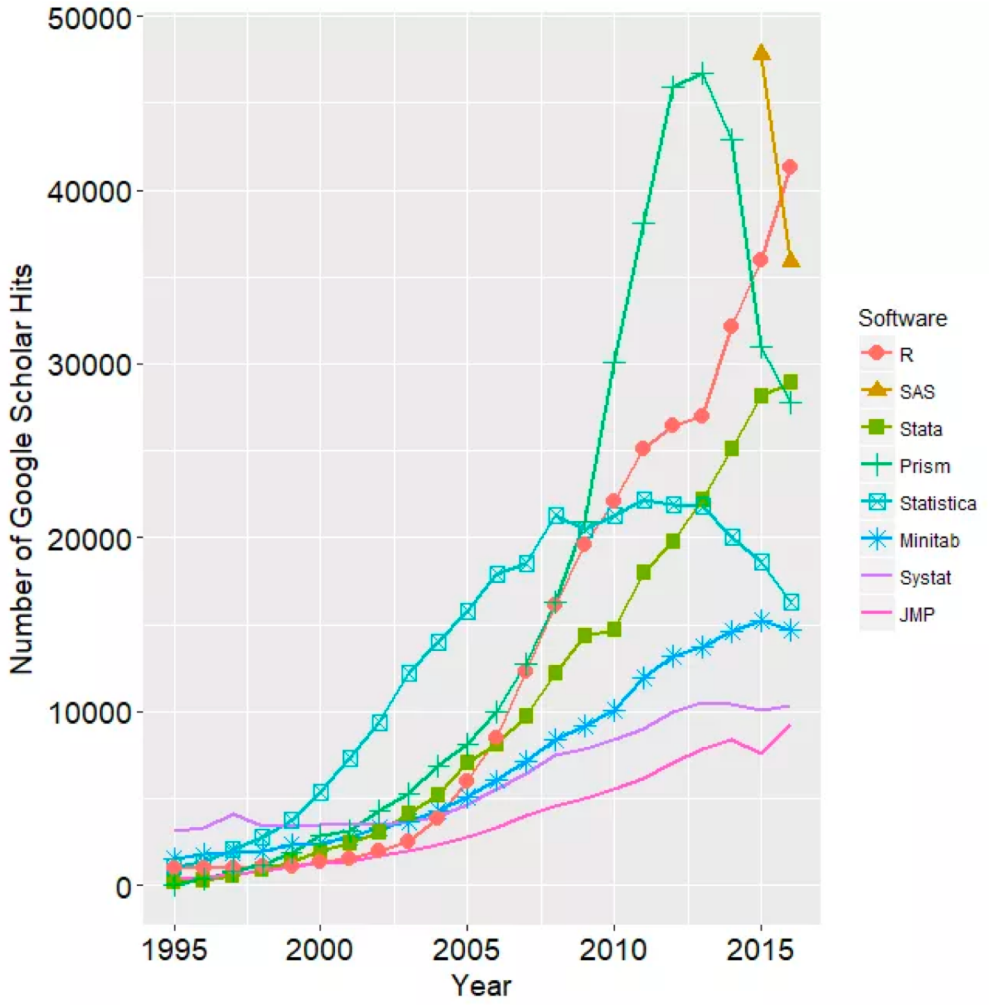
Number of scholarly articles found in the most recent complete year (2016) for each software package used as a topic or tool of analysis. For methods see here.

http://r4stats.com/articles/popularity/

Change in scholarly articles



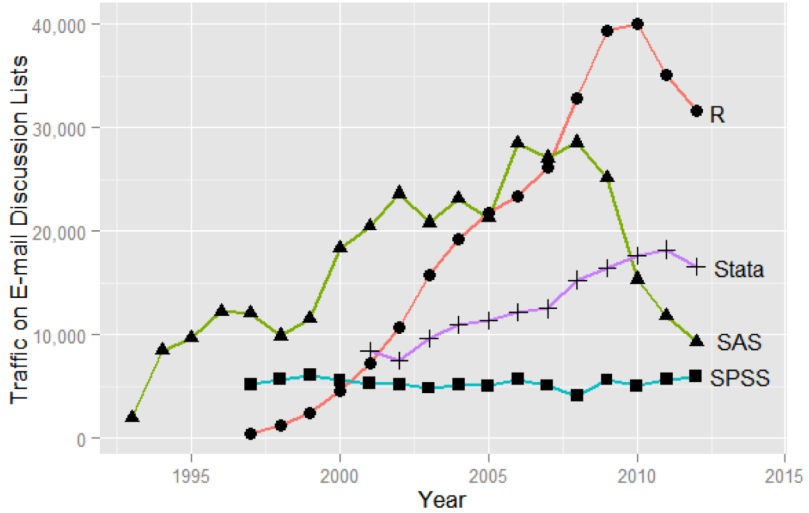
The number of scholarly articles found in each year by Google Scholar. Only the top six “classic” statistics packages are shown.



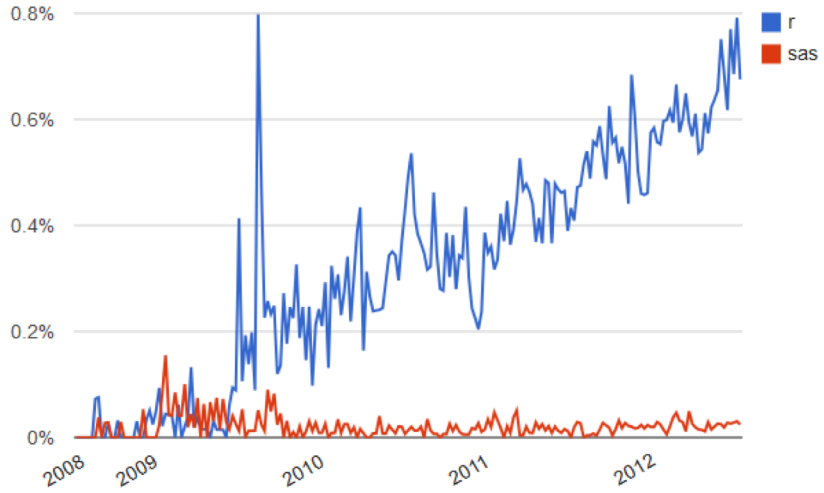
The number of scholarly articles found in each year by Google Scholar (excluding SAS and SPSS).

http://r4stats.com/articles/popularity/

Forum/discussion activity



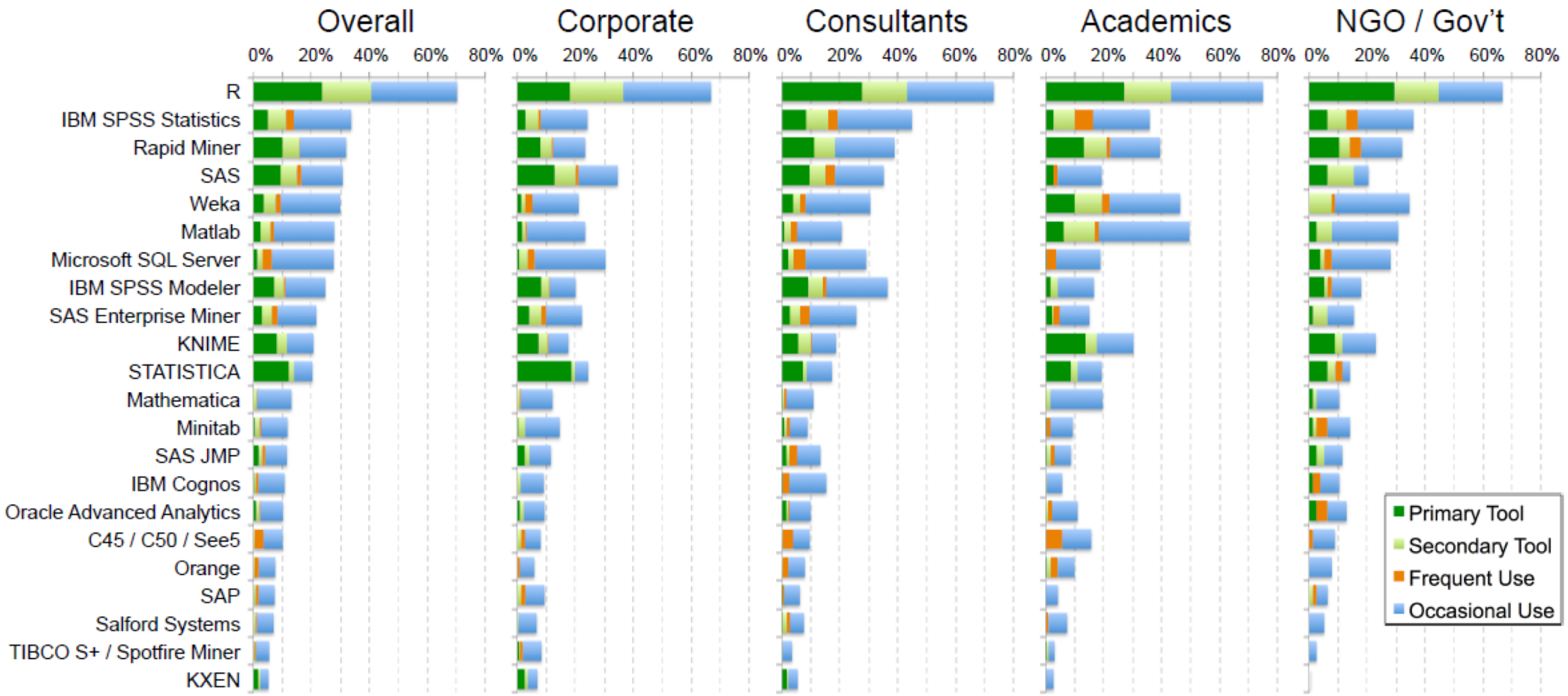
Sum of monthly email traffic on each software's main listserv discussion list.



Number of R- or SAS-related posts to Stack Overflow (programming and statistical topics) by week.

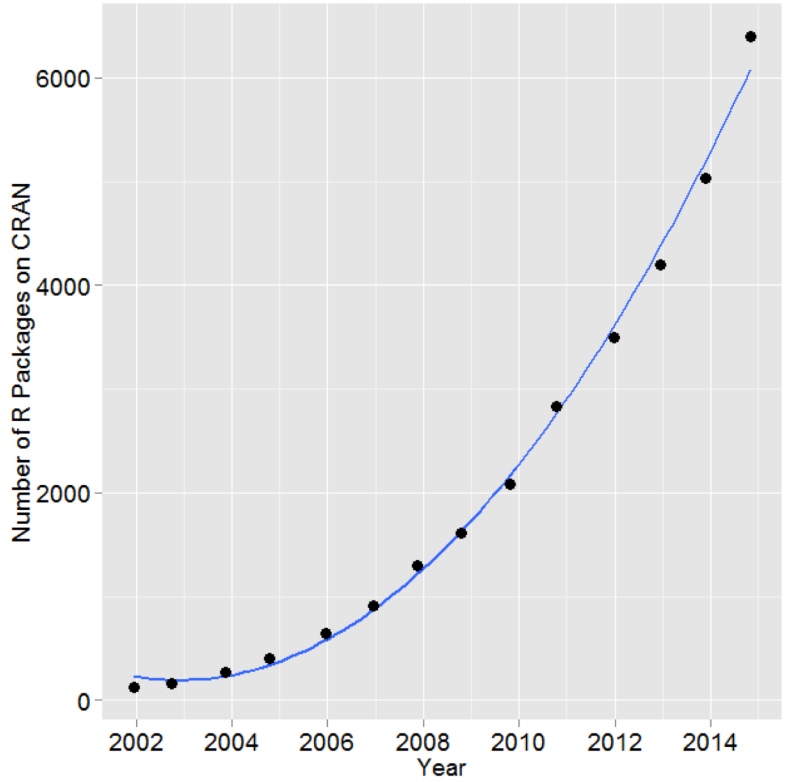
http://r4stats.com/articles/popularity/

Rexer Analytics *Data Miner* Survey (2013)



~1.2k respondents http://r4stats.com/articles/popularity/

R Development



Number of R packages available on its main distribution site for the last version released in each year.

SAS v9.3: 1.2k commands
(in Base, Stat, ETS, HP Forecasting, Graph, IML, Macro, OR, QC.)

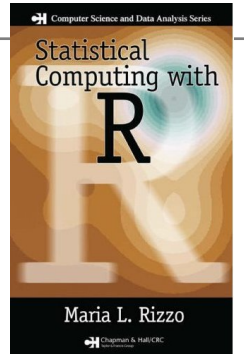
2014: R added 1.3k packages and ~27k functions.

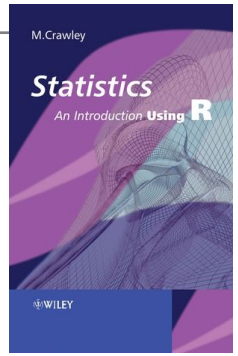
Over 6k packages!

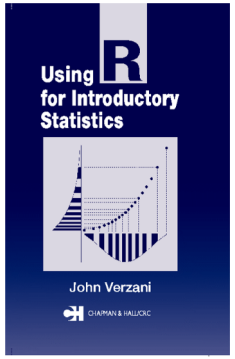
http://r4stats.com/articles/popularity/

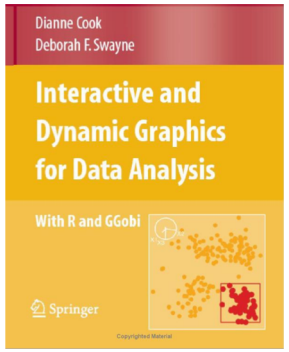
Task Views organize packages by topic: http://cran.r-project.org/web/views/

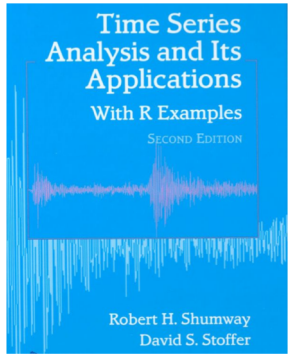
240 Books on R since 2000

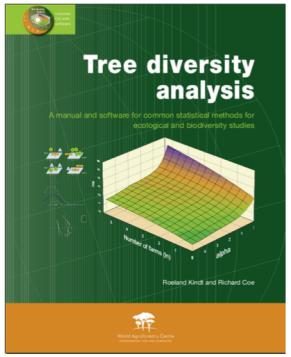


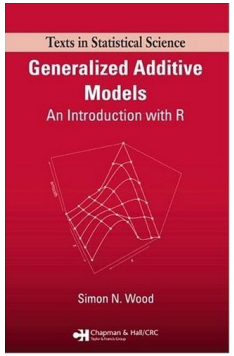


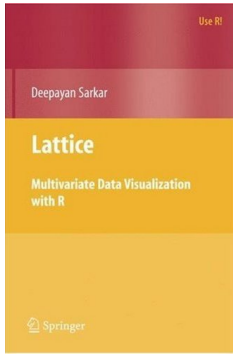


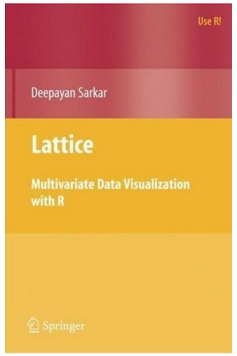


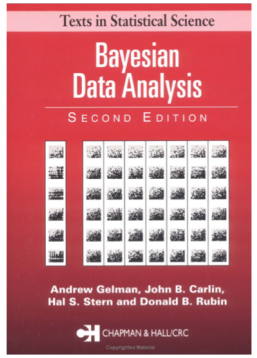


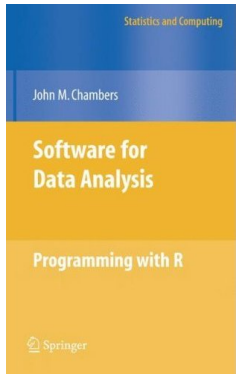


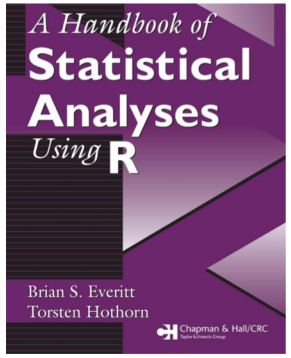


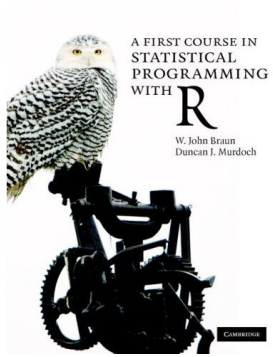


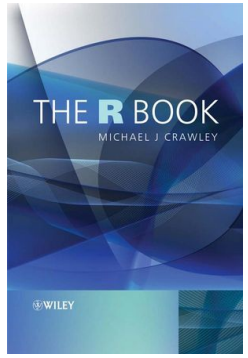


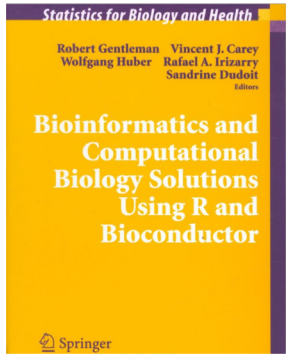


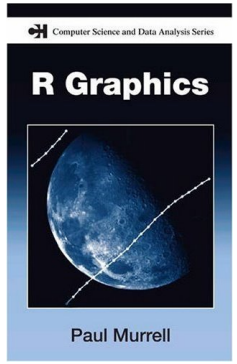


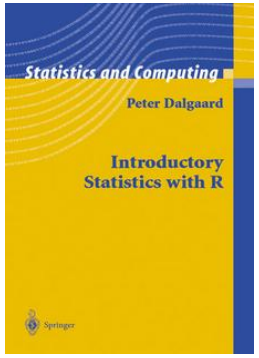


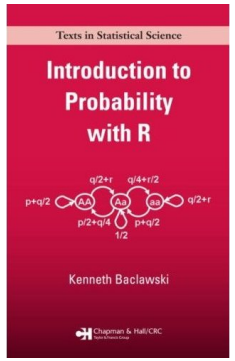








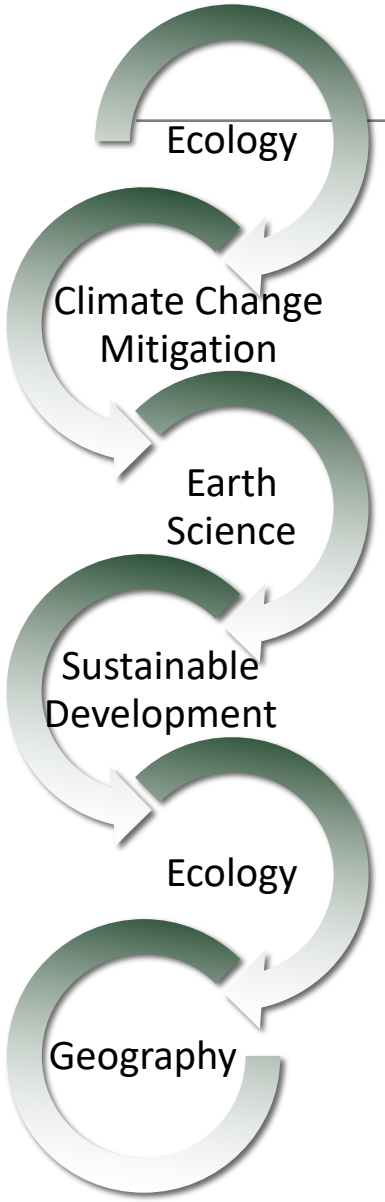




R is a great language to learn and it can take you far…

But others (python, etc.) are useful too. Once you learn one, it is much faster to learn others

A little about me…



GRASS

python

netCDF

Google earth engine
a google.org project

git

NCO

RStudio

NASA Earth Exchange

And who are you?

1. Name
2. Where are you from (state and/or country)?
3. Department/Degree (e.g. MS GIS)
4. Research Interests
5. Motivation for taking this course (what do you want to learn?)

~1 minute each!

Before next class

1. Explore adamwilson.us/SpatialDataScience
2. Install RStudio on your laptop from https://www.rstudio.com
3. Read and work through the first two chapters of *R For Data Science* (link on website)

Installing R and RStudio

Install

- R https://cran.revolutionanalytics.com/
- RStudio https://www.rstudio.com